Introducción a la simulación de datos en Psicología

II Jornada de Metodología Cuantitativa en Psicología - AMP

Brian Norman Peña Calero

Avances en Medición Psicológica

16/10/2020

Acerca de esta presentación

Las diapositivas fueron expuestas en la II Jornada de Metodología Cuantitativa en Psicología organizada por **Avances en Medición Psicológica**.

El video de la ponencia pueden encontrarlo dándole click en el siguiente enlace: **https://www.facebook.com/amp.unmsm/videos/364404511344708/**

Las diapositivas fueron elaboradas mediante el paquete **xaringan** en R 4.0.2. Para una óptima visualización del mismo, recomiendo ir al siguiente enlace: **https://brianmsm.github.io/jornada-amp-simulacion/**, además que podrá siempre tener la versión actualizada de la misma.

El código fuente está disponible en el siguiente enlace: **https://github.com/brianmsm/jornada-amp-simulacion**.

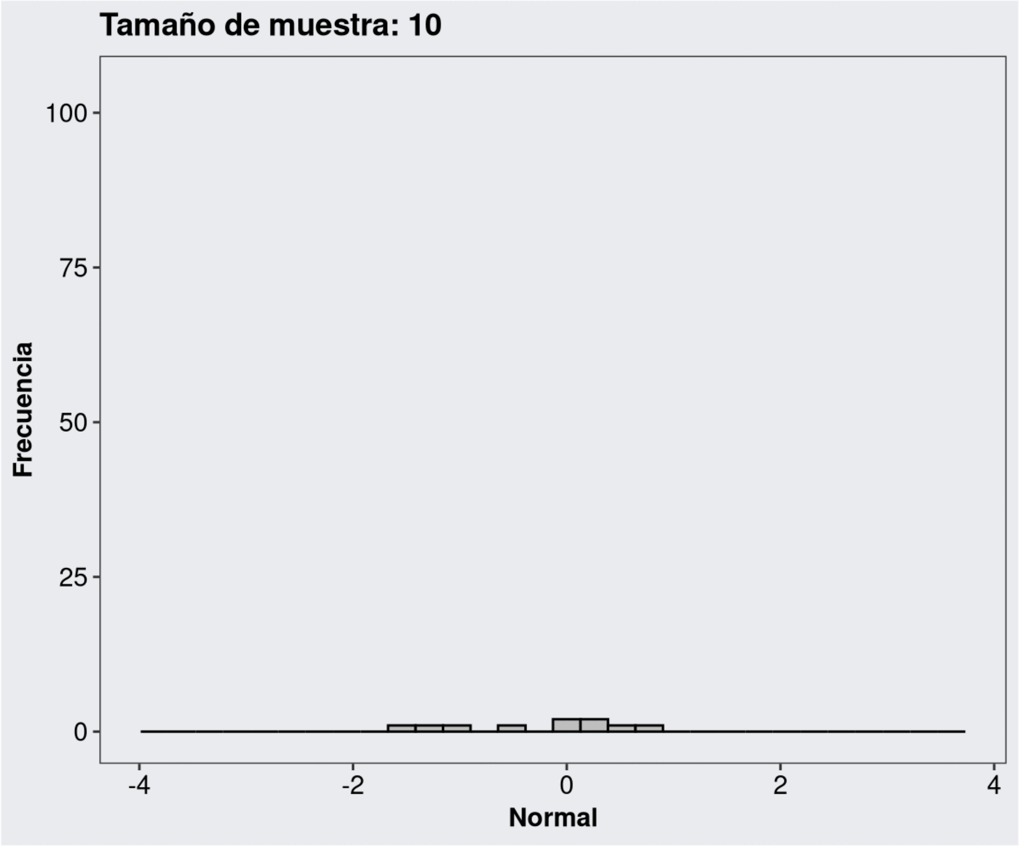
Método Montecarlo

*Es el uso de procesos aleatorios para cuantificar y estudiar distribuciones aleatorias, y a partir de ello analizar y comparar procedimientos estadísticos así como comportamientos de datos en sistemas más complejos (**Gentle, 2005**).*

Para ello necesitamos:

AVANCES EN MEDICIÓN PSICOLÓGICA

- Software para simulación
- Obtener aleatoriamente datos
- Distribución o fórmula subyacente a esa intención
- Condiciones
- Replicaciones
- Evaluación



Software simulación

Existen diversos softwares de costo para realizar simulación de datos (**goldsim**, **xlstat**, **vose**, etc.), sin embargo tienen la limitación de restringirse a funciones y soluciones específicas de determinado sector de interés. Por ej. riesgos financieros.

A fin de tener control total acerca de lo que se hará con los datos, condiciones y formas de evaluarlo, es recomendable utilizar un lenguaje de programación, entre los cuales puede estar C, C++, Ruby, Python, R, etc.

Aunque R, es el lenguaje *más lento* entre los mencionados, es el más difundido en cuanto a análisis de datos (en un sentido similar con `python`) y de fácil entendimiento.

Aleatoriedad de datos

AVANCES EN MEDICIÓN PSICOLÓGICA

Generar un número aleatorio es extremadamente complejo, y conlleva una serie de dificultades y requisitos que no solo se restrinjen al software (**Park and Miller, 1988**), sino también al hardware.

Lo que obtenemos en el software son `números pseudo-aleatorios` puesto que parten de un mismo puerto (`semilla`) pre-determinada para generarse.

```
round(runif(n = 2, min = 1, max = 5), 1) # Pr
```

```
## [1] 2.7 2.7
```

```
round(runif(n = 2, min = 1, max = 5), 1) # Se
```

```
## [1] 1.1 2.5
```

```
set.seed(123) # Establecer semilla
round(runif(n = 2, min = 1, max = 5), 1) # Pr
```

```
## [1] 2.2 4.2
```

```
set.seed(123)
round(runif(n = 2, min = 1, max = 5), 1) # Se
```

```
## [1] 2.2 4.2
```

Distribución o fórmula subyacente

Función de densidad para $X \sim \mathcal{N}(\mu, \sigma^2)$.

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

Función de distribución para $X \sim \mathcal{N}(\mu, \sigma^2)$.

$$F(x) = P(X \leq x)$$

```
x <- seq(-3, 3, 1)
dens_norm <- dnorm(x, mean = 0, sd = 1)
dens_norm
```

```
## [1] 0.004431848 0.053990967 0.241970725 0.3989422
```

```
pnorm(3, mean = 0, sd = 1)
```

```
## [1] 0.9986501
```

```
pnorm(1.96, mean = 0, sd = 1)
```

```
## [1] 0.9750021
```

```
qnorm(0.975, mean = 0, sd = 1)
```

```
## [1] 1.959964
```

Distribución o fórmula subyacente

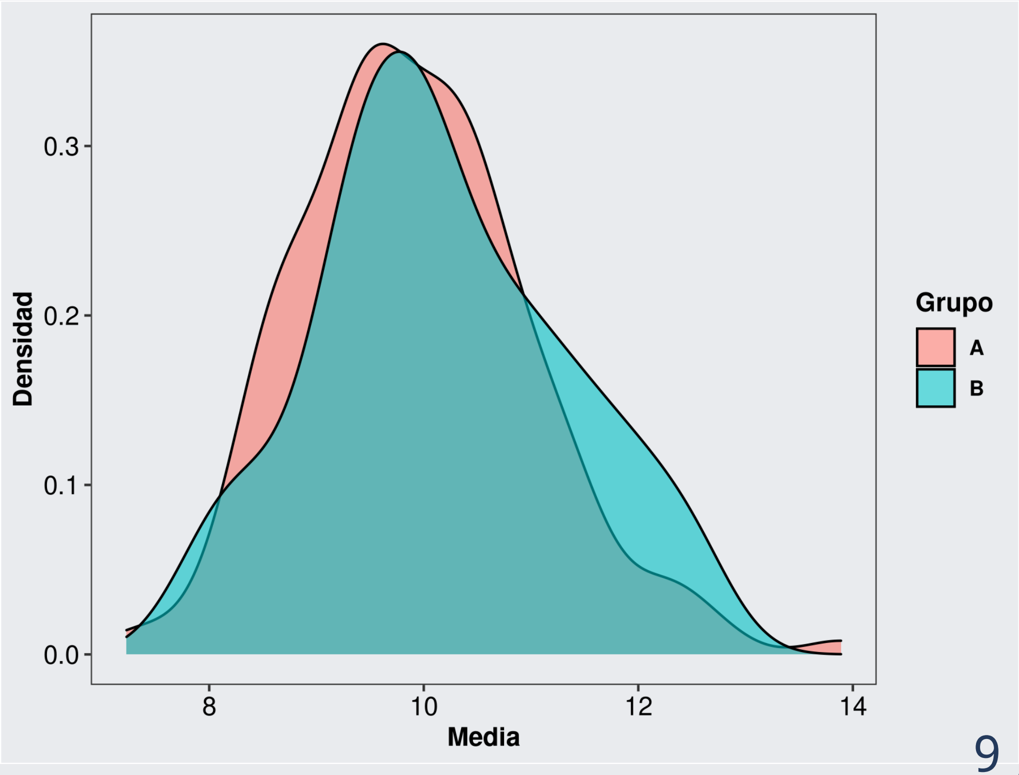
Con la función `rnorm()` se pueden generar números pseudo-aleatorios provenientes de una distribución normal. En base a esto se pueden realizar algunos ensayos.

```
library(tidyverse)
set.seed(123)
dist_normal <- tibble(Media = rnorm(n = 300,
                                    mean = 10,
                                    sd = 1.2))

grupo_a <- dist_normal %>%
  sample_frac(size = 0.5)

grupo_b <- dist_normal %>%
  anti_join(grupo_a)

dist_normal <- bind_rows(grupo_a,
                         grupo_b) %>%
  mutate(
    Grupo = c(rep("A", 150),
              rep("B", 150))
  )
```



Distribución o fórmula subyacente

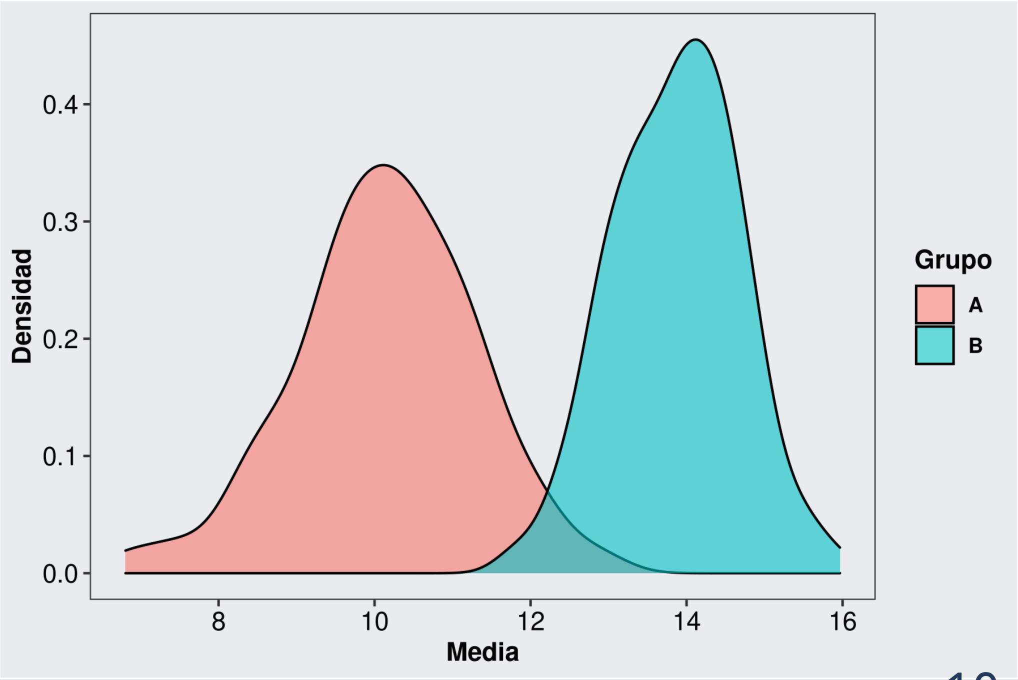
AVANCES EN MEDICIÓN PSICOLÓGICA

Ahora, el ensayo tiene la misma intención pero generando datos aleatorios desde 2 distribuciones distintas. Esto es lo que estaría evaluándose bajo la hipótesis de que ***hay diferencia de medias en 2 grupos***, que en esencia indica que los grupos provienen de distribuciones distintas.

```
dist_norm_A <- tibble(Media = rnorm(n = 150,
                                    mean = 10,
                                    sd = 1.2))

dist_norm_B <- tibble(Media = rnorm(n = 150,
                                    mean = 14,
                                    sd = 0.8))

dist_normal_dif <- bind_rows(dist_norm_A,
                             dist_norm_B) %>%
  mutate(
    Grupo = c(rep("A", 150),
              rep("B", 150))
  )
```



Condiciones

Las condiciones son uno de los componentes claves en el desarrollo de un experimento de simulación Montecarlo, puesto que permite controlar que variables influirán en las evaluaciones y generación de datos que se realicen, y hasta que punto la simulación puede acercarse a condiciones realistas.

Algunas condiciones comunes y de utilidad son:

- Diferenciar tamaños de muestras
- Distribución no-normal de los datos
- Presencia de outliers
- Grupos desiguales
- Presencia de heterocedasticidad

Condiciones

Así, en caso haya 4 variaciones de tamaño de muestra, 3 grupos desiguales, y consideración de normalidad y no-normalidad, se tendría un diseño de 4x3x2 en la generación de los datos.

En ese sentido, si se generá un mínimo de 50 datos por condición, se podría estar trabajando con 3600 datos aproximadamente. Es común observar estudios de simulación que consideren más de 150 condiciones combinadas al mismo tiempo.

En estudios psicométricos, las condiciones pueden aumentar si se toma en cuenta la `cantidad de ítems`, `carga factorial`, `cantidad de factores` y `fiabilidad en cada factor`.

Replicaciones

Las replicaciones permiten controlar problemas de error por la generación de números pseudo-aleatorios. Se trabaja bajo el supuesto de **tendencia**. Si 1 sola generación de datos indica que la prueba de `shapìro-wilk` no detecta adecuadamente la distribución normal de los datos, podría deberse a un error aleatorio.

Sin embargo, la cuestión cambia si se trata de 459 de 1000 veces que se hace la generación de datos exactamente en las mismas condiciones. Los números de replicaciones más comunes oscilan entre 500 y 1000. Por lo que, en el ejemplo anterior tenemos, `4x3x2x1000`, lo que lleva al trabajo con `3 600 000` aproximadamente.

Evaluación

Posterior a la elaboración del diseño y la generación de datos aleatorios correctamente gestionados, se debe evaluar el objetivo del mismo:

- Comportamiento del estimador en diversas condiciones: RMSE y sesgo (**Harwell, 2019**)
- % de Error tipo I y II
- Potencia estadística presente

$$\mathrm{ARB} = \left[\frac{1}{R} \sum_{i=1}^{R} \left(\frac{\hat{\theta} - \theta_i}{\theta} \right) \right] \times 100 \,,$$

$$\mathrm{RMSE}(\mathrm{ARB}) = \frac{1}{R} \sum_{i=1}^{R} \left(\frac{\hat{\theta}_i - \theta}{\theta} \right)^2$$

Aplicaciones:

- **Aprendizaje de estadística**
- **Investigación Metodológica**
  - Análisis del funcionamiento de estadísticos en investigación **empírica**. Ej:
    - Uso de t-student
    - Correlación de Pearson
  - Análisis del funcionamiento de estadísticos en investigación **psicométrica**. Ej:
    - Análisis factorial confirmatorio
    - Coeficiente omega y alfa
    - Índices de ajuste: CFI, TLI, RMSEA, SRMR
  - Funcionamiento de estadísticos en diferentes **condiciones**. Ej:
    - Tamaño de muestras distintos
    - Presencia de no-normalidad
    - Presencia de outliers
    - Data missing

Aprendizaje de estadística

AVANCES EN MEDICIÓN PSICOLÓGICA

Articles

Simulation Methods for Teaching Sampling Distributions: Should Hands-on Activities Precede the Computer?

Stacey A. Hancock & Wendy Rummerfield

Pages 9-17 | Accepted author version posted online: 31 Jan 2020, Published online: 28 Feb 2020

Download citation https://doi.org/10.1080/10691898.2020.1720551 Check for updates

Abstract

Sampling distributions are fundamental to an understanding of statistical inference, yet research shows that students in introductory statistics courses tend to have multiple misconceptions of this important concept. A common instructional method used to address these misconceptions is computer simulation, often preceded by hands-on simulation activities. However, the results on computer simulation activities' effects on student understanding of sampling distributions, and if hands-on simulation activities are necessary, are mixed. In this article, we describe an empirical intervention study in which each of eight

Investigación metodológica

Original Articles

Comparisons of various types of normality tests

B. W. Yap & C. H. Sim

Pages 2141-2155 | Received 03 Jun 2010, Accepted 29 Aug 2010, Published online: 18 May 2011

Download citation https://doi.org/10.1080/00949655.2010.520163

Abstract

Normality tests can be classified into tests based on chi-squared, moments, empirical distribution, spacings, regression and correlation and other special tests. This paper studies and compares the power of eight selected normality tests: the Shapiro–Wilk test, the Kolmogorov–Smirnov test, the Lilliefors test, the Cramer–von Mises test, the Anderson–Darling test, the D'Agostino–Pearson test, the Jarque–Bera test and chi-squared test. Power comparisons of these eight tests were obtained via the Monte Carlo simulation of sample data generated from alternative distributions that follow symmetric short-tailed, symmetric long-tailed and asymmetric distributions. Our simulation results show that for symmetric short-tailed distributions, D'Agostino and Shapiro–Wilk tests have better power. For symmetric long-

Investigación metodológica

AVANCES EN MEDICIÓN PSICOLÓGICA

Original Articles

Minimum Sample Size Recommendations for Conducting Factor Analyses

Daniel J. Mundfrom, Dale G. Shaw & Tian Lu Ke

Pages 159-168 | Published online: 13 Nov 2009

Download citation https://doi.org/10.1207/s15327574ijt0502_4

Abstract

There is no shortage of recommendations regarding the appropriate sample size to use when conducting a factor analysis. Suggested minimums for sample size include from 3 to 20 times the number of variables and absolute ranges from 100 to over 1,000. For the most part, there is little empirical evidence to support these recommendations. This simulation study addressed minimum sample size requirements for 180 different population conditions that varied in the number of factors, the number of variables per factor, and the level of communality. Congruence coefficients were calculated to assess the agreement between population solutions and sample solutions generated from the various population conditions. Although

Fuente: **Tandfonline**

Investigación metodológica

AVANCES EN MEDICIÓN PSICOLÓGICA

Journal of Statistical Modeling and Analytics Vol.2 No.1, 21-33, 2011

Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests

Nornadiah Mohd Razali[1]
Yap Bee Wah[1]
[1]*Faculty of Computer and Mathematical Sciences,*
Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
E-mail: nornadiah@tmsk.uitm.edu.my, yapbeewah@salam.uitm.edu.my

ABSTRACT

The importance of normal distribution is undeniable since it is an underlying assumption of many statistical procedures such as t-tests, linear regression analysis, discriminant analysis and Analysis of Variance (ANOVA). When the normality assumption is violated, interpretation and inferences may not be reliable or valid. The three common procedures in assessing whether a random sample of independent observations of size n come from a population with a normal distribution are: graphical methods (histograms, boxplots, Q-Q-plots), numerical methods (skewness and kurtosis indices) and formal normality tests. This paper compares the power of four formal tests of normality: Shapiro-Wilk (SW) test, Kolmogorov-Smirnov (KS) test, Lilliefors (LF) test and Anderson-Darling (AD) test. Power comparisons of these four tests were obtained via Monte Carlo simulation of sample data generated*

¡Gracias por su atención!